Published online 7 May 2008 *Nucleic Acids Research, 2008, Vol. 36, Web Server issue* ***W509–W512***
doi:10.1093/nar/gkn202

NetMHC-3.0: accurate web accessible predictions of human, mouse and monkey MHC class I affinities for peptides of length 8–11

Claus Lundegaard[1,*], Kasper Lamberth[2], Mikkel Harndahl[2], Søren Buus[2], Ole Lund[1] and Morten Nielsen[1]

[1]CBS, Department of Systems Biology, Technical University of Denmark DTU, Kemitorvet Build. 208, 2800 Lyngby and [2]Department of International Health, Immunology and Microbiology, University of Copenhagen, Panum Institute 22.3.6, Blegdamsvej 18, 2200 Copenhagen N, Denmark

Received January 31, 2008; Revised March 27, 2008; Accepted April 4, 2008

ABSTRACT

NetMHC-3.0 is trained on a large number of quantitative peptide data using both affinity data from the Immune Epitope Database and Analysis Resource (IEDB) and elution data from SYFPEITHI. The method generates high-accuracy predictions of major histocompatibility complex (MHC): peptide binding. The predictions are based on artificial neural networks trained on data from 55 MHC alleles (43 Human and 12 non-human), and position-specific scoring matrices (PSSMs) for additional 67 HLA alleles. As only the MHC class I prediction server is available, predictions are possible for peptides of length 8–11 for all 122 alleles. artificial neural network predictions are given as actual IC_{50} values whereas PSSM predictions are given as a log-odds likelihood scores. The output is optionally available as download for easy post-processing. The training method underlying the server is the best available, and has been used to predict possible MHC-binding peptides in a series of pathogen viral proteomes including SARS, Influenza and HIV, resulting in an average of 75–80% confirmed MHC binders. Here, the performance is further validated and benchmarked using a large set of newly published affinity data, non-redundant to the training set. The server is free of use and available at: http://www.cbs.dtu.dk/services/NetMHC.

INTRODUCTION

Intracellular infections with pathogens such as viruses and certain bacteria are defeated by cytotoxic T lymphocytes (CTL). The CTL T-cell receptor (TCR) recognizes foreign peptides in complex with major histocompatibility complex (MHC) class I molecules on the surface of the infected cells. MHC class I molecules preferably bind and present nine amino acid long peptides, which mainly originates from proteins expressed in the cytosol of the presenting cell. In most vertebrates, MHCs exist in a number of different allelic variants that each binds a specific and very limited set of peptides. For a number of years, prediction methods have developed to identify which peptides will bind a given MHC (1), and such predictions can be highly valuable in a broad range of applications, including rational vaccine design and disease diagnostics. The artificial neural network (ANN) training method behind NetMHC (2,3) has been benchmarked to be the best among available methods (4). Preliminary versions of the algorithm have been used to predict possible MHC-binding peptides in a large set of pathogenic viral proteomes, resulting in an average of >75% confirmed MHC binders (5). Most MHC prediction algorithms (a list of other servers is included in the Supplementary Material) are trained on peptides of the same length as they predict, but since data for peptide lengths different from nine are much more scarce, the broadness of MHC binding predictions for different peptide lengths is accordingly limited. In this server, however, a method is implemented making it possible to predict 8-, 10- and 11-mer peptide binding using 9-mer trained predictors, which extends the MHC coverage for these peptide lengths significantly compared to other available MHC:peptide-binding servers.

METHODS

The server is trained on the largest number of quantitative peptide:MHC affinity measurements ever published using both affinity data from the Immune Epitope Database and Analysis Resource (IEDB) (6), eluted peptide data from

*To whom correspondence should be addressed. Tel: +45 21900767; Fax: +45 45931585; Email: lunde@cbc.dtu.dk

the SYFPEITHI database (7) and proprietary affinity data. The predictions based on ANNs are trained essentially as described in (3) on data from 55 MHC alleles (43 Human and 12 non-human), and the predictions based on position specific scoring matrices (PSSMs) are trained as described in (2) for additional 67 HLA alleles. A large number of 9-mer MHC affinity data have become available from the IEDB database, since the training of the ANNs used at NetMHC-3.0, and all peptides not used in the training (6452 9-mer peptide affinity data points, covering 32 HLA alleles) were used for evaluation of the server performance. These data are available at the server. In this dataset, 3104 were measured to be binders (IC_{50}<500 nM), 76% of these were correctly predicted as such. 3030 peptides were predicted to bind to a given HLA, and 78% of these had a measured IC_{50}<500 nM. The average Pearson correlation coefficient (PCC) and area under a ROC curve (AUC) value using a 500 nM classification threshold were 0.71 and 0.86, respectively. For the full per allele results, see the Supplementary Material (Supplementary Table 1 and Supplementary Figure 1). NetMHC-3.0 uses a new approximation algorithm that reliably predicts the affinity of peptides of lengths 8, 10 and 11, for which affinity data for training are rare (8). The method uses predictors trained on peptides of length 9 to successfully extrapolate to other lengths. In short, the method approximates each peptide of any length to a number of 9-mers, by inserting X (for 8-mers) or deleting amino acid(s) (for 10- and 11-mers) and set the final prediction to an average of the 9-mer predictions. We had previously trained ANN predictors directly on 10-mer affinity data and since this training more than 2000 10-mer peptide:MHC affinities had become available from the IEDB database (6). Area under a ROC curve (AUC) values were calculated for each allele using either ANNs trained on 10-mers or the approximation method. For 12 of the 16 alleles, the approximation method performed better than the 10-mer trained ANNs ($P < 0.01$), see Supplementary Material Figure 2. However, for the four HLA-alleles, this evaluation showed better performance for ANNs trained on 10-mer peptides; these 10-mer trained ANNs are used for predictions by the server. For 8-mers, 2002 affinity data were extracted covering 35 MHC alleles. The overall PCC and AUC were 0.68 and 0.86, respectively. For 8-mer per allele performance, see the Supplementary Material Figure 4. For 8-mers, predictors trained on actual 8-mers seems to be better than the approximation method otherwise used, so for the alleles with available 8-mer affinity data, 8-mer trained ANNs are used for the predictions. In general, it is not possible to estimate how reliable a single prediction is. However, the stronger the affinity is predicted the higher are the chance that the actual affinity is stronger than the generally accepted binding threshold of 500 nM.

SERVER

NetMHC-3.0 predicts the binding affinity of either a list of peptides with a defined length (8–11 residues) or all possible sub-peptides hosted within full-length proteins.

```
>Protein_1
SLYNTVATLSLYNTATLSLYNTVATL
>Protein_2
ELEVENTENNINEANDEIGHTMERS
```

Figure 1. Example input in FASTA format.

The input must be in the FASTA format, or as peptides all of equal length, one peptide pr. line. The server will accept a maximum of 5000 sequences per submission; each sequence not more than 20 000 amino acids with a minimum length corresponding to the selected length of prediction (see subsequently). Input data can be pasted into a text field or uploaded from a local file on the user's computer.

If the input is in peptide, format the corresponding tick-box must be selected. The input must not exceed 5000 sequences and with a maximum of 20 000 amino acids in each sequence. One or more MHCs must be selected, as well as the desired peptide length. Only one prediction length at a time can be used. The output can optionally be sorted according to the predicted affinity by selecting a tick-box. The predictions start by clicking the Submit button. An example input in FASTA format is shown in Figure 1.

The output is displayed as raw text with a header indicating the server name, the type of prediction (PSSM, ANN or ANN-approximation) the first selected allele and the date (Figure 2) followed by the prediction output in a column format. The columns are named in the first line of the prediction output. The first column [pos] is the position of the first amino acid of the predicted peptide within the possibly longer sequence, numbering starting with 0. Column (peptide) is the primary sequence of the (sub-)peptide. Column (logscore) is the raw prediction output, which for ANNs is $1\text{-}\log_{50000}$ to the affinity in nanomolar units. For PSSM predictions the raw prediction score is a log-odds likelihood score. Additionally a column is included for ANN predictions, [affinity (nM)], which is the predicted affinity presented in nanomolar units. Column (Bind Level) indicates if the peptide is predicted to bind stronger than a certain threshold [for ANN predictions stronger than 50 nM (SB) or stronger than 500 nM (WB); for PSSM high-binding peptides (SB) have a prediction score greater than the 0.1% percentile score value of 1 000 000 random natural peptides, and weak binding (WB) peptides a score value above the 1% percentile score of 1 000 000 random natural peptides predictions]. Predicted affinities weaker than 500 nM or lower than the 1% percentile score have no indications. Column (Protein Name) gives the name of the predicted protein. If peptide input was used, the name will always be 'Sequence'. Column (Allele) gives the name of the MHC allele chosen. The output contains all the sub-peptides for each protein for a given allele either in the order they appear in the sequence or sorted by predicted affinity within each protein (if chosen). If more than one protein sequence were entered, a dashed line will separate the

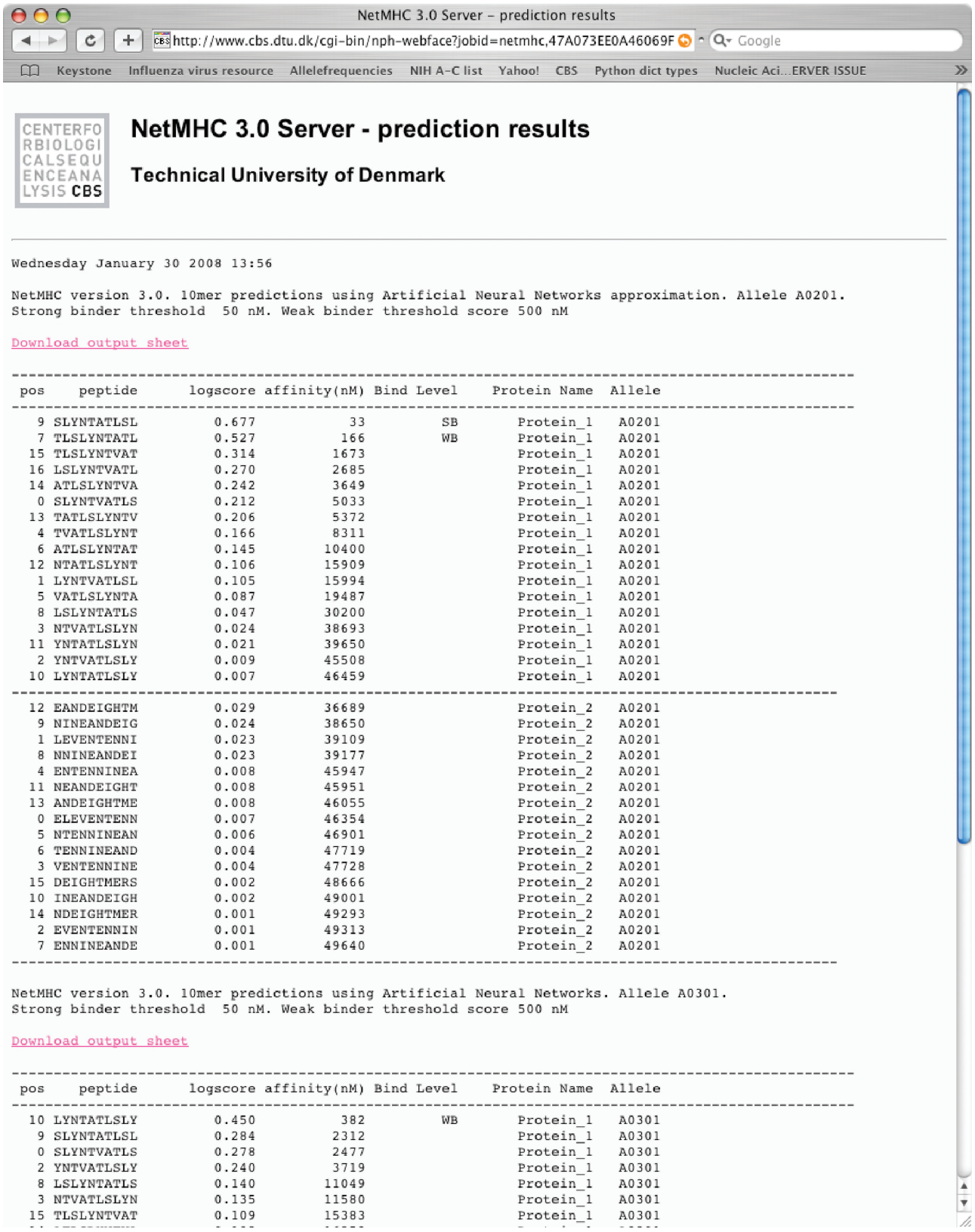
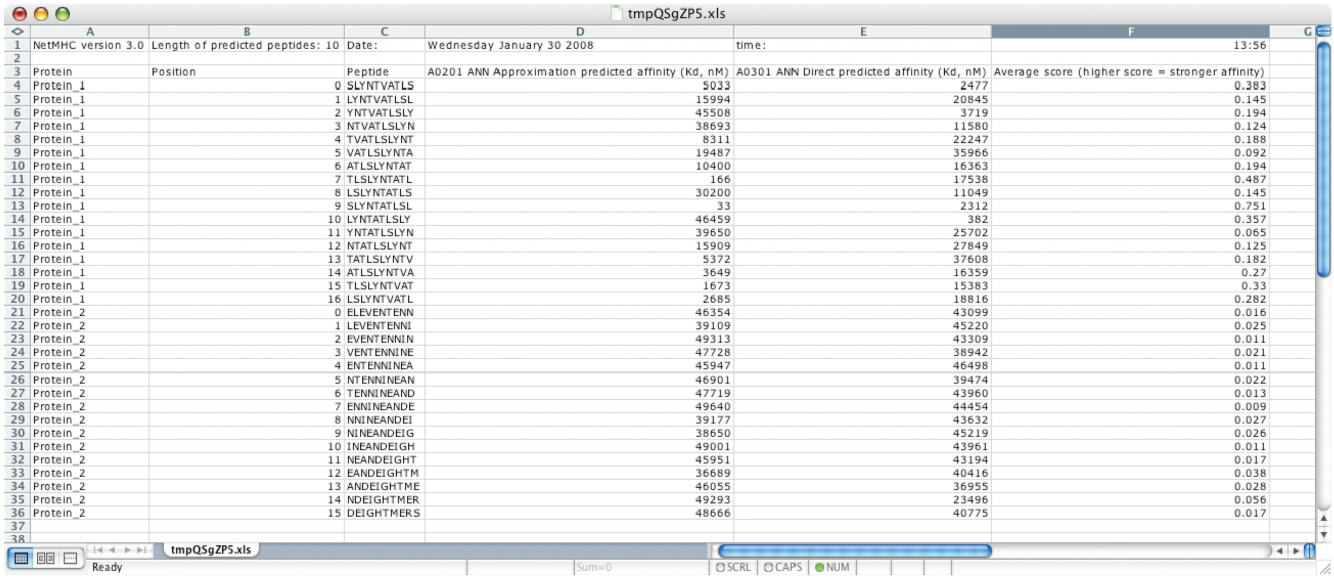


Figure 2. Raw text output using the input in Figure 1 and selecting the alleles HLA-A0201 and HLA-A0301. 10-mer peptide predictions were chosen. Affinity sorting was chosen.



tmpQSgZP5.xls

	A	B	C	D	E	F	G
1	NetMHC version 3.0	Length of predicted peptides: 10	Date:	Wednesday January 30 2008	time:	13:56	
2							
3	Protein	Position	Peptide	A0201 ANN Approximation predicted affinity (Kd, nM)	A0301 ANN Direct predicted affinity (Kd, nM)	Average score (higher score = stronger affinity)	
4	Protein_1	0	SLYNTVATLS	5033	2477	0.383	
5	Protein_1	1	LYNTVATLSL	15994	20845	0.145	
6	Protein_1	2	YNTVATLSLY	45508	3719	0.194	
7	Protein_1	3	NTVATLSLYN	38693	11580	0.124	
8	Protein_1	4	TVATLSLYNT	8311	22247	0.188	
9	Protein_1	5	VATLSLYNTA	19487	35966	0.092	
10	Protein_1	6	ATLSLYNTAT	10400	16363	0.194	
11	Protein_1	7	TLSLYNTATL	166	17538	0.487	
12	Protein_1	8	LSLYNTATLS	30200	11049	0.145	
13	Protein_1	9	SLYNTATLSL	33	2312	0.751	
14	Protein_1	10	LYNTATLSLY	46459	382	0.357	
15	Protein_1	11	YNTATLSLYN	39650	25702	0.065	
16	Protein_1	12	NTATLSLYNT	15909	27849	0.125	
17	Protein_1	13	TATLSLYNTV	5372	37608	0.182	
18	Protein_1	14	ATLSLYNTVA	3649	16359	0.27	
19	Protein_1	15	TLSLYNTVAT	1673	15383	0.33	
20	Protein_1	16	LSLYNTVATL	2685	18816	0.282	
21	Protein_2	0	ELEVENTENN	46354	43099	0.016	
22	Protein_2	1	LEVENTENNI	39109	45220	0.025	
23	Protein_2	2	EVENTENNIN	49313	43309	0.011	
24	Protein_2	3	VENTENNINE	47728	38942	0.021	
25	Protein_2	4	ENTENNINEA	45947	46498	0.011	
26	Protein_2	5	NTENNINEAN	46901	39474	0.022	
27	Protein_2	6	TENNINEAND	47719	43960	0.013	
28	Protein_2	7	ENNINEANDE	49640	44454	0.009	
29	Protein_2	8	NNINEANDEI	39177	43632	0.027	
30	Protein_2	9	NINEANDEIG	38650	45219	0.026	
31	Protein_2	10	INEANDEIGH	49001	43961	0.011	
32	Protein_2	11	NEANDEIGHT	45951	43194	0.017	
33	Protein_2	12	EANDEIGHTM	36689	40416	0.038	
34	Protein_2	13	ANDEIGHTME	46055	36955	0.028	
35	Protein_2	14	NDEIGHTMER	49293	23496	0.056	
36	Protein_2	15	DEIGHTMERS	48666	40775	0.017	
37							
38							

tmpQSgZP5.xls

Ready

Sum=0

SCRL CAPS NUM

Figure 3. Downloaded output sheet opened in Microsoft® Excel and adjusted with of column. The output was generated using input in Figure 1 and selecting the alleles HLA-A0201 and HLA-A0301. 10mer peptide predictions were chosen.

peptides from each protein. If more than one allele were chosen, the output will show a header similar to the first immediately after the first predictions, all in the same web output page.

In each header, there is a link to a file with the output in tab as separated format, where the filename ends on.xls making it easily imported into spreadsheet programs. This file always contains the predicted peptides in the order they appeared in the input file. The output data for each peptide will be displayed on a single line with predictions for each of the selected alleles in different columns (Figure 3).

FINAL REMARKS

This server is developed to aid research and limit the resources needed for rational and effective CTL epitope discovery and will be continuously updated as new data become available. All comments and suggestions for usability improvements are most welcome.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

This work was funded by European Commission (LSHB-CT-2003-503231, LSHB-CT-2004-012175) and National Institutes of Health (HHSNN26600400006C, HHSN266 200400025C, HHSN266200400083C).

Conflict of interest statement. None declared.

REFERENCES

1. Lundegaard,C., Lund,O., Kesmir,C., Brunak,S. and Nielsen,M. (2007) Modeling the adaptive immune system: predictions and simulations. *Bioinformatics*, **23**, 3265–3275.
2. Nielsen,M., Lundegaard,C., Worning,P., Hvid,C.S., Lamberth,K., Buus,S., Brunak,S. and Lund,O. (2004) Improved prediction of MHC class I and class II epitopes using a novel Gibbs sampling approach. *Bioinformatics*, **20**, 1388–1397.
3. Nielsen,M., Lundegaard,C., Worning,P., Lauemoller,S.L., Lamberth,K., Buus,S., Brunak,S. and Lund,O. (2003) Reliable prediction of T-cell epitopes using neural networks with novel sequence representations. *Protein Sci*, **12**, 1007–1017.
4. Peters,B., Bui,H.H., Frankild,S., Nielson,M., Lundegaard,C., Kostem,E., Basch,D., Lamberth,K., Harndahl,M., Fleri,W. *et al.* (2006) A community resource benchmarking predictions of peptide binding to MHC-I molecules. PLoS Comput. Biol., **2**, e65.
5. Sylvester-Hvid,C., Nielsen,M., Lamberth,K., Roder,G., Justesen,S., Lundegaard,C., Worning,P., Thomadsen,H., Lund,O., Brunak,S. *et al.* (2004) SARS CTL vaccine candidates; HLA supertype-, genome-wide scanning and biochemical validation. *Tissue Antigens*, **63**, 395–400.
6. Sette,A., Fleri,W., Peters,B., Sathiamurthy,M., Bui,H.H. and Wilson,S. (2005) A roadmap for the immunomics of category A-C pathogens. *Immunity*, **22**, 155–161.
7. Rammensee,H., Bachmann,J., Emmerich,N.P., Bachor,O.A. and Stevanovic,S. (1999) SYFPEITHI: database for MHC ligands and peptide motifs. *Immunogenetics*, **50**, 213–219.
8. Lundegaard,C., Lund,O. and Nielsen,M. (2008) Accurate approximation method for prediction of class I MHC af-finities for peptides of length 8, 10 and 11 using prediction tools trained on 9mers. *Bioinformatics*, in press, doi:10.1093/bioinformatics/btn128.